文章编号：2095-2163(2020)09-0225-04 中图分类号：TP338.6 文献标志码：A

基于应用场景的人工智能芯片技术分类方法研究

赵春昊
（上海依图网络科技有限公司，上海 200051）

摘 要：AI 芯片通过利用适配人工智能算法的运算架构为 AI 应用提供高性能算力，成为 AI 领域的研究热点。如何从技术角度对不同应用场景的 AI 芯片进行分类是需要填补的研究空白。本文从不同的应用场景，即云侧、边缘侧、端侧出发，从 AI 芯片的技术形态、计算任务、技术领域等角度开展研究，总结了每类芯片的核心需求、技术以及相应的参数特点，制定了一个简单有效的分类判断依据和标准。同时，本文收集整理了市场上不同应用场景的主流 AI 芯片性能参数，并且和上述分类标准进行比较，二者的参数范围得到相互印证，因此证明了本文提出的分类方法的有效性。

关键词：人工智能；芯片；分类方法

A classification method of AI chip based on application scenarios

ZHAO Chunhao
（Shanghai YITU Network Technology Co., Ltd, Shanghai 200051, China）

【Abstract】 AI chips are architectures that can provide high performance computing power for AI applications, which has become a research hotspot in the field of AI. How to classify AI chips in different scenario from a technical perspective has become an important research topic. In this paper, we summarize the technical features of AI chips in different application scenario and formulate a simple and effective classification principle and standard from different application scenarios, i.e. cloud side, edge side and end side. Meanwhile, the performance parameters of mainstream AI chips in different application scenarios in the market are collected and compared with the above classification criteria, and thus proving the effectiveness of the classification method proposed in this paper.

【Key words】 artificial intelligence; chip; taxonomy

0 引 言

近年来，人工智能（AI）已经成为信息技术中最热门的领域之一。随着算法精度要求的不断提升，对于智能计算性能的要求也越来越高。AI 芯片作为新型基础设施的核心，逐渐成为了该领域的研究热点。AI 芯片是指具备适配人工智能算法的运算架构，能够完成人工智能应用运算处理的集成电路元件。由于 AI 应用的底层算子具有通用性，AI 芯片可以利用基础计算单元针对性地提供高性能算力。

AI 芯片不仅仅在云端超算和数据中心里使用，而且逐渐向边缘侧和终端场景发展。例如，英伟达 Xavier NX 芯片作为自动驾驶车载芯片的领跑者，可以完成目标检测、路径规划等 AI 任务的实时计算，这类 AI 芯片已经成为了边缘侧计算的关键部件。而许多小型终端设备也嵌入了 AI 芯片。新出现的边缘侧和端侧计算设备中的 AI 芯片呈现出与云端 AI 芯片完全不同的性能。如何从技术角度对 AI 芯片进行分类的依据成为需要填补的研究空白。

2018 年，张蔚敏等从产业的角度分析了 AI 芯片的市场需求、机遇与挑战，并且按照两个维度（技术架构、市场需求）对于 AI 芯片进行分类[1]；清华大学微电子学研究所尹首一等侧重于计算的角度，分析了现阶段 AI 加速芯片的技术特点，展望了类脑仿生芯片、通用 AI 芯片等未来趋势[2]；丛瑛瑛等分析了 AI 芯片目前的发展态势，提出了对策和建议[3]；美国麻省理工大学的 Albert Reuther 等从性能与功耗角度分析了市面主流的 AI 芯片性能，并且使用性能与功耗对其分类[4]。本文从更细致的角度针对不同应用场景和技术领域内的 AI 芯片性能进行深入的研究，制定一种有效的 AI 芯片分类方法。

1 问题描述

AI 芯片的技术特点，取决于其应用环境的计算需求。随着 AI 芯片的应用场景不断扩大，其越来越呈现出多样性发展的技术特点。各类应用场景对于 AI 芯片的需求是不同的，因此影响了芯片厂商的设计思路，并最终反映为 AI 芯片在性能参数上的差异。如何针对各类业务场景对 AI 芯片的特点进行归纳总结并分类，在不同分类下功能与性能参数的标准如何界定，这些问题成为 AI 芯片研究领域内亟

作者简介：赵春昊（1982-），男，硕士，中级工程师，主要研究方向：人工智能。

收稿日期：2020-07-03

需解答的问题。

2 各应用场景下 AI 芯片的分类

本文讨论的 AI 芯片包括 GPU(图形处理单元),NPU(神经处理单元)类脑芯片和 SoC(片上系统)以及 FPGA(可编程门阵列),这些芯片的定义都可以在前人的工作中找到。需要注意的是,虽然 CPU(中央处理单元)可以执行传统机器学习算法,但由于其算力较小且不适合处理大规模数据集,故严格意义上并不属于 AI 芯片范畴。

AI 芯片的应用场景可以分为“云侧”、“边缘侧”与“端侧”。云侧场景是指大规模集中式和并发处理计算任务的场景;边缘侧场景是指利用网络的边缘节点来汇聚、处理、分析多个终端设备数据的场景;而端侧场景则是设备直接与用户、环境交互,主要进行数据获取及本地处理功能的场景。

AI 芯片需要根据不同应用场景的需求,结合芯片技术特点进行设计。同时,由于计算任务的不同,对算力的要求也不尽相同。不同的 AI 技术领域所产生的数据类型和数据量大小差异巨大。前人的工作主要是对 AI 芯片的技术形态进行了界定与简单分析。但由于不同应用场景下,AI 芯片的形态差别巨大,因此本文在此基础上,从不同的应用场景,即云侧、边缘侧、端侧入手,从 AI 芯片的技术形态、计算任务、技术领域等角度开展研究,并试图进行芯片的技术参数分类,制定一个简单有效的分类判断依据和标准。

2.1 按 AI 芯片技术形态分类

在不同应用场景下,有不同类型的典型设备,其加载的 AI 芯片技术形态也不同。云侧对应的是大型集中式计算集群;边缘侧有自动驾驶系统、机器人、边缘网关等;而端侧设备则包括智能移动终端、可穿戴设备、传感器等。通过对各种芯片应用场景的梳理发现,端侧 AI 设备基本都以 SoC 形态出现;边缘侧主要以小规模的 SoC 独立运行,或者以装配加速卡的服务器形态出现,而云侧主要是以带加速卡的服务器组成大规模集群,或者大规模 SoC 集群形态出现。

不同应用场景下 AI 芯片的技术形态见表 1。

表 1 不同应用场景下 AI 芯片的技术形态

Tab. 1 Technical forms of AI chips in different application scenarios

技术形态	云侧	边缘侧	端侧
加速卡(GPU,NPU)	√(多机集群)	√(单机运行)	
微处理器式 SoC			√(单机运行)
全功能式 SoC	√(多机集群)	√(单机运行)	√(单机运行)
FPGA	√	√	√

2.2 按 AI 芯片计算任务分类

人工智能的计算任务可以分为训练与推理。有一些 AI 芯片在设计之初就定为专门负责训练任务或推理任务,这些芯片一般被称为训练芯片或推理芯片(对于加速卡而言,又被称为训练卡或加速卡)。

训练芯片和推理芯片呈现出完全不同的技术参数特点。对于训练芯片而言,由于训练过程需要大量的算力,需要处理的数据量也较大,因此所有的训练芯片都在云侧。而这些训练芯片大都可以在不同的技术领域通用。推理任务需要利用训练后的模型,针对应用场景下的输入,实时给出计算结果。随着 AI 计算应用场景的拓展,推理芯片适用场景也不断拓宽,在云侧、边缘侧和端侧都有使用。各应用场景下,不同计算任务的 AI 芯片典型实例见表 2。

表 2 各类计算任务在不同应用场景下的典型 AI 芯片

Tab. 2 Typical training/inferencing AI chips in different application scenarios

计算任务	云侧	边缘侧	端侧
训练芯片	P100/V100/A100(英伟达) 昇腾 910(华为海思) 含光 800(阿里平头哥)	/	/
推理芯片	T4(英伟达) 思元 270-S4(寒武纪)	Xavier(英伟达) 昇腾 310(华为海思)	旭日系列(地平线) 玄铁 910(阿里平头哥) A13 Bionic(苹果)

2.3　按 AI 芯片技术领域分类

AI 芯片可以划分为通用 AI 芯片和专用 AI 芯片。通用 AI 芯片是指可以支持多个细分领域的 AI 芯片；而专用 AI 芯片指的是只为某个特定细分行业服务的 AI 芯片。

在专用 AI 芯片中，根据 AI 芯片计算任务所属的技术领域，总结了“视觉分析”、“自动驾驶”、“智能语音”、“路径规划”、“金融风控”、“辅助诊断”这六个人工智能目前最重要的落地领域[5]，收集了其典型设备参数，并对核心需求进行了总结，见表 3。

表 3　各类技术领域在不同应用场景下的典型设备、核心需求、技术关键总结

Tab. 3　A summary on typical devices, core requirements and key techniques in different application scenarios in each technical field

技术领域	典型设备			核心需求			技术关键		
	云	边	端	云	边	端	云	边	端
视觉分析	AI 服务器	边缘服务器	智能摄像头	大算力	高带宽 准确率 高算力	实时性	算力	带宽 算法 算力	带宽 延时
自动驾驶		Xavier板卡	DCU TI智能仪器		超高带宽 准确率 高算力	安全性 超低延时		带宽 算法 算力	延时
智能语音		智能语音集成系统	声纹采集 实时转译 语音助手		/	准确率 实时性		/	算法 延时
路径规划		/	车载 GPS		/	准确率 中延时		/	算法 延时
金融风控		风控检测	高频交易计算设备		低延时 高召回率	超低延时 高准确率		延时 算法	延时 算法
辅助诊断		院级数据平台	终端医疗		储存量	准确率		数据量	算法

可以发现：视觉分析、自动驾驶、金融风控这三大领域对 AI 芯片算力、延时、安全性等要求最高，也恰恰是专用 AI 芯片出现最多的领域，进一步证明了 AI 芯片的设计是为了满足技术领域下的各应用场景服务的。

3　基于应用场景的 AI 芯片分类标准

AI 芯片的参数与其应用需求密切相关，从每个应用场景下的芯片需求出发，考虑各场景下 AI 芯片的核心参数，并以此建立分类标准。收集国内外主流芯片参数来验证这一分类标准的可行性。

3.1　各应用场景 AI 芯片参数需求

(1)云侧芯片。云侧设备往往不需要满足实时性，又支持大规模扩展。因此，云侧设备对于功耗没有限制，同时会在较大程度上追求有效的算力。云侧 AI 芯片的功耗大多都超过 100 W，而 AI 芯片的能效比大多在 1 TOPS/W 以上，因此芯片的算力基本都大于 100 TOPS。另外，目前大多数云侧 AI 加速卡都是用 PCIe3.0 连接，而其功耗限制为 300 W，大多数 AI 芯片功耗都没有超过这个限制。

(2)边缘侧芯片。由于要实时完成本地数据汇聚和计算，因此边缘侧设备的 AI 芯片的参数需求是围绕其本地业务需求设计的，各技术领域内的参数差异巨大。例如，对于视觉分析，一个边缘视频盒子一般需要接入 4-20 路的视频，因此其芯片需要 20-100 MB/s 的带宽来支撑视频流的输入，而其算力则需要根据其计算的特定任务，如：人脸检测、特征提取、匹配等来决定，目前大约在 1-1.5 TOPS/路，因此边缘盒子的算力需求在 5-30 TOPS 左右，而对应的功耗大约在 5-50 W。自动驾驶领域，技术参数性能要求更高，由于自动驾驶需要同时接收和处理多个传感器的数据，因此芯片带宽大约需要 10 GB/s 左右。而根据英伟达预测，其 L2 层级(部分自动化驾驶)的算力和功耗需求大约分别为 200 TOPS 和 80 W[6]。而相对而言，自然语言处理、语音分析等领域内的算力要求不高。

(3)端侧芯片。由于端侧设备体积小，且常常需要在无线状态下使用电池供电，因此功耗成为重要的参数要求。例如：手机，由于其芯片散热面积有限，导致其功耗受限。市面上几乎所有的手机芯片的散热设计功耗都在 5 瓦以下。随着 5G 等技术的发展，更多的计算要求可以被加载到更大型的边缘侧计算设备上，故普通场景下的端侧推理需求对算力的要求不高。例如，被誉为有“智能手机有史以来最好的机器学习性能”的苹果 A12 芯片，其算力也仅仅在 5 TOPS。

不同应用场景的技术参数需求估计，见表 4。

表 4 不同应用场景的技术参数需求估计

Tab. 4 Estimations on chip specifications in different application scenarios

云侧		边缘侧			端侧
单颗芯片	算力:>100TOPS	计算机视觉领域	L2 自动驾驶:	语音分析:	算力:0.1-5TOPS
典型参数	功耗:100-300 W	算力:5-30TOPS	算力:200TOPS	算力:10TOPS	功耗:<5W
		功耗:5-50W	功耗:80 W	功耗:10W	
		带宽:20-100 MB/s	带宽:10 GB/s		

3.2 国内、国际主流芯片的参数分析

进一步收集市面上主流的芯片,选取了其中有代表性的芯片来验证上述分类标准。按照不同芯片的应用场景划分类,按照参数峰值功耗和算力这两个维度进行统计,结果如图 1 所示。

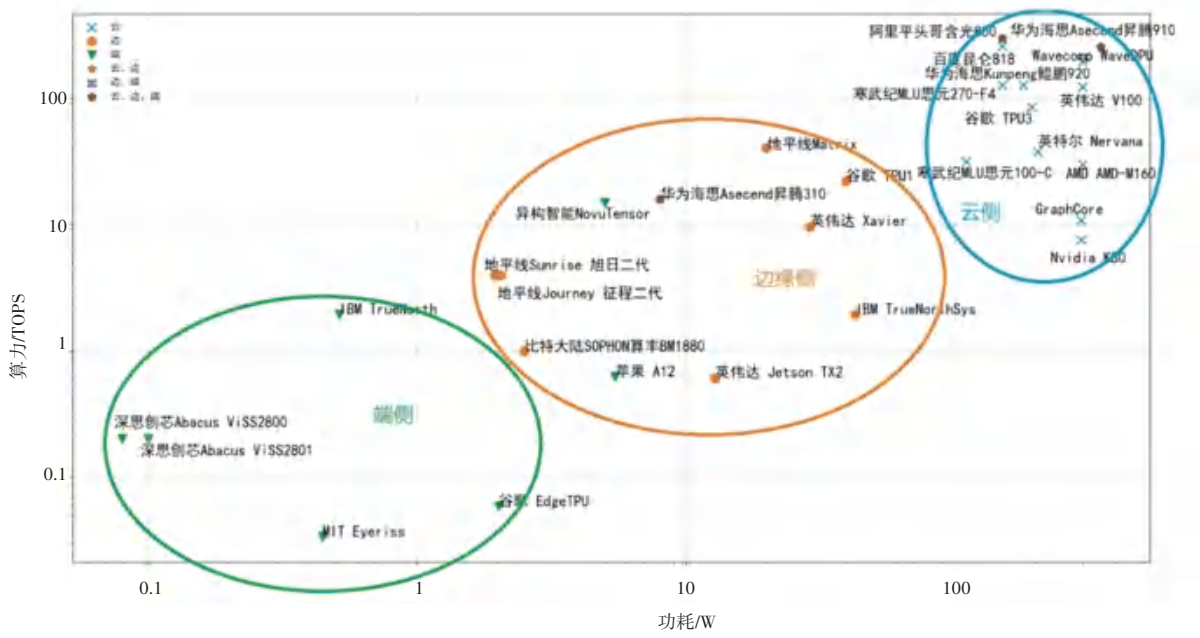

图 1 国内外不同应用场景典型芯片的参数统计

Fig. 1 Statistics on specifications of typical chips in different application scenarios

根据图 1 中信息显示,可以总结出典型 AI 芯片在不同应用场景下的分类标准:

(1)端侧 AI 芯片:算力<5 TOPS,功耗<5 W。

(2)边缘侧 AI 芯片:算力 5-100 TOPS,功耗 5-100 W。

(3)云侧 AI 芯片:算力>100 TOPS,功耗 100 W-300 W。

结果与需求角度分析的结果得到了相互印证。

4 结束语

本文研究了 AI 芯片在技术参数上的特点与其在各个应用场景下分类依据。首先从“技术形态”,“计算任务”,“技术领域”三个维度进行分类,总结了每种类别下芯片的典型产品,核心能力与芯片需求。从需求角度分析了不同应用场景 AI 芯片所需要的性能参数,得出了云侧、边缘侧和端侧芯片的分类标准。研究发现可以从功耗和算力这两个角度联合分析出不同应用场景下的芯片共同点。云侧芯片的功耗大约是 100 W-300 W,端侧芯片的功耗大多在 5 W 以下,而边缘侧芯片则介于两者之间;而从算力角度来看,划分的界线大致为 5 TOPS(端侧-边缘侧)和 100 TOPS(边缘侧-云侧)。进一步收集了市面主流芯片的参数,发现目前不同场景的芯片参数符合本文提出的分类标准,验证了本文所提出的分类标准的有效性。

参考文献

[1] 张蔚敏, 蒋阿芳, 纪学毅. 人工智能芯片产业现状[J]. 信息通信技术与政策, 2018.

[2]尹首一, 郭珩, 魏少军. 人工智能芯片发展的现状及趋势[J]. 科技导报, 2018, 36(17):45-51.

[3]丛瑛瑛, 陈丝. 人工智能芯片发展态势分析及对策建议[J]. 信息通信技术与政策, 2018 (8):65-68.

[4]Reuther A, Michaleas P, Jones M, et al. Survey and benchmarking of machine learning accelerators[J]. arXiv preprint arXiv:1908.11348, 2019.

[5] Deloitte. Semiconductors-the Next Wave[R]. 2019.

[6] NVIDIA. GTC 2020KEYNOTE[Z/OL]. 2020.